DOCUMENT RESUME

ED 389 719 TM 024 219

AUTHOR Morgan, Rick; Mazzeo, John
TITLE A Comparison of the Structural Relationships among Reading, Listening, Writing, and Speaking Components of the AP French Language Examination for AP Candidates and College Students.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-59
PUB DATE Oct 88
NOTE 56p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Advanced Placement Programs; *College Students; Comparative Analysis; Error of Measurement; Experience; Factor Structure; *French; Higher Education; High Schools; *High School Students; Language Proficiency; *Listening; Multiple Choice Tests; *Reading Comprehension; Second Language Learning; Speech; Test Items; Writing (Composition)
IDENTIFIERS *Advanced Placement Examinations (CEEB); Confirmatory Factor Analysis; Exploratory Factor Analysis; Structural Constituents

ABSTRACT

The dimensional structure of the 1987 Advanced Placement (AP) French language examination was tested in four populations using a series of confirmatory linear factor analysis models. To mitigate problems with the linear factor analysis of multiple choice items, the linear factor analysis of item parcel scores, made of small mutually exclusive collections of items hypothesized to measure the same underlying dimension, was used. Six confirmatory factor analysis models were tested with each of five data samples. Two contained high school AP candidates with no out-of-school French language experience (n=1,500 each), and a third consisted of candidates who had spent a significant amount of time in a French-speaking country (n=1,418). A fourth sample contained 477 AP candidates who regularly spoke or heard French at home, and the final sample contained 302 students with no out-of-class experience enrolled in third year college French. In all samples the examination appeared to measure four major dimensions associated with listening, reading, writing, and speaking. For those without out-of-school experience, the examination displayed invariance of factor loadings and errors of measurement. Factor structures were similar for groups with similar out-of-school French language experience. (Contains 15 tables, 1 figure, 2 tabular appendixes, and 13 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 719

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

A COMPARISON OF THE STRUCTURAL RELATIONSHIPS AMONG READING, LISTENING, WRITING, AND SPEAKING COMPONENTS OF THE AP FRENCH LANGUAGE EXAMINATION FOR AP CANDIDATES AND COLLEGE STUDENTS

Rick Morgan
John Mazzeo

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
October 1988

TM024219

A Comparison of the Structural Relationships among Reading, Listening, Writing, and Speaking Components of the AP French Language Examination for AP Candidates and College Students

Rick Morgan and John Mazzeo

Copyright © 1988. Educational Testing Service. All rights reserved.

Abstract

Yearly the Advanced Placement Program administers an examination to high school students measuring French language skills that students might acquire after six semesters of college French Language courses. The dimensional structure of the 1987 AP French Language exam was tested in four populations using a series of confirmatory linear factor analysis models. In order to mitigate problems associated with the linear factor analysis of multiple-choice items, the linear factor analysis of item parcel scores, made up of small mutually exclusive collections of items hypothesized to measure the same underlying dimension, was utilized. Six confirmatory factor analysis models were tested within each of five samples of data. Two samples contained high school AP candidates with no out-of-school French Language experience. A third sample contained AP candiates which had spent a significant amount of time in a French speaking counry. A fourth sample contained AP candidates who regularly spoke or heard French at home. The final sample contained students with no out-of-class French language experience enrolled in third year French classes at one of sixteen colleges. In all samples the exam appears to measure four major dimensions which are associated with the language skills of listening, reading, writing and speaking. For the student groups lacking out-of-school French language experience, the structure of the exam displays invariance of factor loadings and errors of measurement. Factor structures were most similar for groups with similar out-of-school French language experiences.

A Comparison of the Structural Relationships among Reading, Listening, Writing, and Speaking Components of the AP French Language Examination for AP Candidates and College Students[1,2]

Rick Morgan and John Mazzeo
Educational Testing Service

In May of each year, the Advanced Placement Program (AP Program) administers an examination in French Language to students, most of which are enrolled in a corresponding high school Advanced Placement course. The exam is intended to be a measure of the listening, reading, writing, and speaking skills that one might acquire after completing six semesters of college French Language courses (College Board, 1987). The majority of the examinees have acquired their knowledge of French through secondary school study. The exam is intended primarily as a measure of French proficiency for this group of examinees. However, some examinees have spent significant time in a French speaking country or either speak or hear French in the home. Ideally, the test is also intended to be an appropriate measure for these latter two groups.

The AP French Language examination consists of 100 4-option multiple-choice questions and a 65 minute free-response section. The multiple-choice and free-response sections are divided into subsections as follows:

Multiple-Choice

Items 1-20	Listening 1 - Reply to a remark or question
Items 21-40	Listening 2 - Questions following monologue or dialogue passages
Items 41-60	Language structure
Items 61-100	Reading comprehension

Free-response

Section 1	Short answer fill-in (0-10 points)
Section 2	Short answer fill-in (0-10 points)
Section 3	Writing essay (0-9 points)
Section 4	Speaking - Directed Responses (0-24 points)
Section 5	Speaking - Picture Story (0-9 points)

---

[1]This research was supported by the College Board Program Funds

[2]The authors wish to thank Daniel Eignor and Rebecca Zwick for their reviews of this paper.

The multiple-choice section begins with 40 items covering two types of listening tasks. The first 20 listening items require the examinee to reply to a remark or question, while the second set of 20 listening items are based upon four longer monologue or dialogue passages. These initial 40 questions are intended to measure listening skills. Twenty language structure items covering grammar and structure follow. A 40 question reading comprehension section completes the multiple-choice section. The latter 60 items (i.e., the language structure and reading comprehension sections) are intended to be measures of reading skills. The multiple-choice section is formula scored by taking the number of correct answers minus one-third the number of incorrect answers. Omitted items receive no points.

The free-response section is intended to measure the abilities of examinees to write and speak in French. The writing section contains two sets of ten fill-in questions and an essay task. The fill-in questions are divided into two sets of ten questions each. For each of these sets, each of the ten fill-in questions are scored as correct (1 point) or incorrect (0 points) by a single grader. Thus, each examinee receives a pair of scores on the fill-in items (i.e., a score for the first ten items and a score for the second ten items), each of which can range from 0 to 10. The essay question is scored by a single grader, using a 1 to 9 scale, with a score of 0 reserved for no response.

The speaking portion of the free-response section consists of two tasks. The first speaking task (directed responses) requires the examinee to verbally respond to each of six questions. The second task (picture story) requires the examinee to view a sequence of pictures illustrating a story and then to interpret and verbally describe the picture story. Examinee responses to both these tasks are tape-recorded and the tapes are sent to ETS where they are scored as part of the normal AP essay grading. Each question in the first task is scored by a single grader using a 1 to 4 scale, with 0 reserved for no response. Thus, each examinee receives a score of 0 to 24 for the directed responses task. The picture story task is scored by a single grader using a 1 to 9 scale, with 0 reserved for no response.

Each examinee receives a total composite score on the exam which consists of a weighted sum of the multiple-choice and free-response components. The weights used in forming the composite scores are derived

such that, ideally, the four language skills (i.e., listening and reading, which are measured in the multiple-choice section, and writing and speaking, which are measured in the free-response section) each contribute 25% to the composite score.

The division of the AP French Language exam into listening, reading, writing, and speaking sections has been done on the basis of the test development committee's notions as to the knowledge and skills measured by the various item types. However, prior to the present study, there had been no formal investigation of the internal structure of the test to determine whether that structure was consistent with prior notions of the knowledge and skills measured by the exam. Futhermore, a lack of information exists as to how similar the internal structure of the exam is for students with varying French language backgrounds. Thus, the primary goal of this research is to obtain a fuller understanding of skills being measured by the AP French Language examination and to determine whether the structure of the test is constant across populations. More specifically, this study examines the internal covariance structure of the AP French Language exam to determine if that structure is consistent with prior notions as to the number and types of skills being measured.

Applying factor analysis models to a test like the AP French Language exam is less than straightforward due to the inclusion of both multiple-choice and free-response questions. The five free-response tasks each resulted in scores which can take on a number of possible score categories (0 to 9 for the essay and picture story tasks, 0 to 10 for the two sets of fill-in items, and 0 to 24 for the directed responses). Linear factor analysis models can be applied to such scores since it is not unreasonable to postulate linear relations between latent factors and observed variables which can take on a number of values, provided that the normality assumptions inherent in such models are not severely violated.

In contrast, multiple-choice items are not appropriately modeled by linear factor analytic techniques, since such items result in scores with either two categories (correct/incorrect) or three categories (correct/omit/incorrect), depending on whether the items are formula scored. With such a small number of score categories, assumptions concerning linear relationships between factors and observed variables, as well as normality

assumptions, are violated (MacDonald & Ahlawat, 1974, Mislevy, 1987). Carroll (1983), Cook, Dorans, and Eignor (1988), Dorans and Lawrence (1987), MacDonald and Ahlawat (1974), Mislevy (1987), and Zwick (1986) have all discussed the problems inherent in the linear factor analysis of such data using matrices of phi coefficients. In brief, the analysis produces spurious "difficulty factors" which are misleading as to the number of content or skill dimensions which might underlie the observed data.

A variety of alternative methods have been tried for the factor analysis of binary-scored data. These are reviewed in some detail in a number of articles (Cook, Dorans, & Eignor, (1988); Dorans & Lawrence (1987); Mislevy (1987); Zwick (1986)). One such method involves estimating and analyzing a matrix of tetrachoric correlations, corrected for guessing, in lieu of the phi matrix. Dorans and Lawrence (1987) obtained results which suggest that analyzing a matrix of corrected tetrachoric correlations among multiple-choice items does not produce satisfactory results in that difficulty factors were not completely eliminated. An additional practical problem with estimating tetrachoric correlations is that the resulting matrix is often not positive definite, prohibiting the use of generalized least squares or maximum likelihood methods for estimating the parameters in the factor analysis models.

Two promising approaches in the factor analysis of binary data are demonstrated by the work of Bock and colleagues involving full information factor analysis (Bock, Gibbons & Muraki, 1986) and the work of MacDonald and colleagues in the area of nonlinear factor analysis (MacDonald & Ahlawat, 1974). Unfortunately, the full information approach, as operationalized in TESTFACT (Wilson, Wood, & Gibbons, 1984), was too costly to be applied to a test with a multiple-choice section the size of the French Language exam as well as not being suited to an exam with essay and free-response components. For the nonlinear factor analysis approaches, conflicting results have been reported as to its utility, thus no attempt was made to apply this procedure in the present study. (See Hambleton & Rovinelli, 1986, for a successful use and Hattie, 1984 for some concerns.)

Given the statistical and practical problems involved in the linear factor analysis of multiple-choice item data and the expense of TESTFACT, the approach utilized in this study involved the linear factor analysis of item

parcel scores made up of small mutually exclusive collections of items hypothesized to measure the same underlying dimension. Such an approach has been used by Cook, Dorans, and Eignor (1988) and Dorans and Lawrence (1987) in factor analytic studies of the SAT. In the parcel approach, scores on individual items are not used directly in deriving covariance matrices for factor analysis. Instead, covariances among scores on sets of related items (parcels) are obtained.

The purpose of the parcel approach is to avoid the statistical problems associated with nonlinear item/factor regressions and scores that are not normally distributed by creating parcel scores, each of which can take on a number of possible score values. One attempts to create parcels so as to linearize the regression of observed variables on the factors and to avoid producing scores with distributions that are extremely skewed.

The parcel approach proceeds in the following manner. First, one identifies groups of items which are intended to measure some common attribute, skill, or content area. Each group of items is then divided into some number of parcels, each of which is made of several items within the group. To the extent possible, the parcel score, as measured by the percent of the maximum score in the parcel, is created to have approximately equal means and standard deviations as other parcels within the group. In addition, the production of extremely difficult or easy sets of items is avoided to circumvent, as much as possible, problems associated with non-normality.

METHOD

The data analyzed in this study were obtained from the 1987 edition of the AP French Language examination.

Populations and Samples

Factor analyses were performed on data from four different groups of examinees. For one of these groups, two independent samples were drawn. Thus, a total of five sets of factor analyses were carried out. Three of the groups consisted of AP candidates (i.e., high school students) who took the exam in May of 1987 as part of the normal AP administration procedure. The fourth group consisted of college students who were administered the exam in

the spring of 1987 as part of a criterion-related validity study (Morgan & Maneckshana, 1988).

The first group of AP candidates were those with little or no out-of-school French language experience. It is for this group of students, referred to hereafter as the "standard group", that the French Language Examination is primarily intended. In 1987, the test was administered to 5,854 standard group students. For the purposes of this study, two non-overlapping spaced samples of 1500 students each were chosen and are designated as standard group samples 1 and 2. The second sample was selected to allow for cross-validation of the findings obtained from the first sample.

The second group of AP candidates are those whose instruction in French comes primarily from secondary school classes, but who have also spent a significant amount of time in a country in which French is routinely spoken. The 1987 French Language exam was administered to 1418 students who indicated having spent at least one month in a French speaking country. Factor analyses were conducted using data from this entire group of examinees, hereafter referred to as "special group 1".

The third group of students are those who are either native speakers of French or who come from homes in which French is regularly spoken. The 1987 French Language exam was administered to 477 such students who reported regularly speaking or listening to French at home. Factor analyses were conducted using data from this entire group of examinees, herafter referred to "special group 2".

The fourth population consisted of the 302 college students with little or no out-of-school French language experience. The college students completed the entire exam, under motivated circumstances, at one of the 16 institutions listed in the Appendix 1.

Formation of item parcels

The multiple-choice items, within each of the four subsections (i.e., listening 1, listening 2, language structure, and reading comprehension), were separated into item parcels. The complete sets of listening 1 items, listening 2 items, and language structure items were each divided into three parcels. Two parcels from each section contained seven items, and a third parcel contained six items. The set of Reading Comprehension items were divided into six parcels, four of which contained seven items, and two of

which contained six items. Thus, a total of fifteen multiple-choice parcels were produced. As discussed above, the intention was to produce subsets of items which were approximately equivalent in difficulty, within item sections, and which produced parcel score distributions with approximately equal standard deviations.

Following the formation of the multiple-choice item parcels, the parcels were formula scored using a correction for guessing. Formula scores on the parcels were computed for each examinee, by awarding 4 points for correct responses, 1 point for an omit or unreached question and no points for an incorrect response. The scoring used whole numbers rather than fractions because of technical reasons related to the computer program used for scoring the multiple-choice portion of the test.

The free response section did not require parcelling. Scores for each of the five free-response subsections were used directly in the analyses. The means and standard deviations (as a percentage of their maximum score) of all 20 parcels (i.e., 15 multiple-choice parcels and 5 free-response section scores) are given in Appendix 2 for each of the five samples for which analyses were performed.

Matrices of the covariances among the 15 multiple-choice parcels and the 5 free-response scores were obtained for each of the five analysis samples. These covariance matricies served as input for linear confirmatory factor analyses. Covariance, rather than correlation, must be analyzed to allow for the testing of various degrees of model invariance across groups, as described below (Joreskog & Sorbom, 1984). The correlation matrices associated with each of these five covariance matrices are given in Appendix 3.

Factor analysis approach

The covariance structure of the test was studied using linear factor-analysis models. The models were of the form:

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E}$$

where: **Y** is a n-by-1 random vector of n observed variables
X is a k-by-1 random vector of k latent variables
B is an n-by-k matrix of coefficients for the linear regression of **Y** on **X**
E is an n-by-1 random vector of residuals,
X and **E** are uncorrelated, and

all variables are expressed as deviations about their means. (1)

Such models imply the following covariance structure:

$$\Phi_y = B(\Gamma_x)B' + \theta_e$$

where: Φ_y denotes the n-by-n covariance matrix of Y
B' is the transponse of the matrix B
Γ_x denotes the k-by-k covariance matrix of X, and
θ_e denotes the n-by-n covariance matrix of the residual variable E (2)

All factor analysis models were estimated using LISREL VI (Joreskog & Sorbom, 1984). The results reported were obtained by maximum likelihood estimation procedures, which make the assumption that the vector variable Y has a multinormal distribution.

Factor Analysis Models

Six different factor models were constructed based on substantive considerations and prior hypotheses as to the possible internal structure of the test. In all six models, the unique components associated with each of the observed variables were assumed to be mutually uncorrelated (i.e., θ_e was assumed to be a diagonal matrix). In addition, for all six models, the correlations between the factors were not constrained to be zero (i.e., Γ_x was a symmetric matrix with nonzero off-diagonal elements). The factor pattern matrices for each of the models, which show the pattern of fixed and free loadings for the B matrices, are presented in Figure 1. For example, the first short listening parcel (L1-1) is allowed to load on the single factor in model M1F, on only one of the two factors in model M2F, and on only one of the factors in models M4FA, M4FB, M5F, and M6F.

The one-factor model (M1F) was generated from a hypothesis that the AP French Language exam might be a unidimensional test, measuring a single language proficiency factor, rather than the multiple proficiencies of listening, reading, writing, and speaking. The two-factor model (M2F) assumes a multiple-choice factor and a free-response factor. This model was generated from a hypothesis that proficiencies with tne question formats (multiple-choice vs free response), rather than different types of language skills, might account for the structure of the data.

Two four factor models were developed. The first of these (M4FA) was developed under the hypothesis that: 1) both the shorter listening items and

the listening items based on longer passages are measures of a common listening factor; 2) the language structure and reading comprehension items measure a common reading factor; 3) the fill-in items and the essays measure a writing factor; and 4) the directed responses and picture story measure a speaking factor. M4FA is based upon the stated design of the test. The second four-factor model (M4FB) also includes factors for listening, speaking, reading, and writing. However, in this model the language structure items are removed from the reading factor and placed with the writing items. Thus, the second four-factor model was generated from a hypothesis that the free-response writing tasks, which are graded in part for correctness of grammar, and the multiple-choice language structure items are measuring a common factor while using different item formats.

In addition, one five factor and one six factor model were developed. The five-factor model (M5F) contains separate listening, language structure, reading comprehension, writing, and speaking factors. The six-factor model (M6F) is an expanded version of the five-factor model which includes separate factors for the two types of listening tasks, those requiring responses to questions or brief remarks (listening section 1) and those requiring responses to longer passages such as monologues and dialogues (listening section 2).

Assessing Model Fit

Following both Joreskog and Sorbom (1984) and Dorans and Lawrence (1987), a descriptive approach to assessing model fit was employed. **LISREL VI** provides a number of fit indices. In the present paper, we evaluated model fit by jointly considering the magnitudes of two of these indices. In addition, model fit was evaluated in light of the the differences between observed and fitted values associated with elements of the covariance matrix being analyzed.

When maximum likelihood estimation of model parameters is used, **LISREL VI** provides a likelihood ratio chi-square statistic, its associated degrees of freedom, and its probability level. Ideally, the likelihood ratio chi-square (for individual models) and differences between chi-square statistics (for pairs of appropriately related models), could be used in a hypothesis testing mode to select a "best fitting" model. However, Joreskog and Sorbom

(1984) caution against such use, pointing out that the distribution of the chi-square statistic is extremely sensitive to departures from multivariate normality in the observed data, as well as being sensitive to sample size. They suggest treating chi-squares (or differences in chi-squares) in a descriptive manner. The size of the chi-square should be evaluated against its associated degrees of freedom. Small values, relative to degrees of freedom, suggest good fitting models. Large differences in the chi-squares, relative to differences in degrees of freedom, for competing models suggest that one of the two models provides substantially better fit to the data.

The second goodness-of-fit index considered is the root mean-squared residual (RMSR). It is defined as:

$$RMSR = [2\ \Sigma_i\Sigma_j\ (s(i,j) - r(i,j))^2 / k(k+1)\ ]^{1/2}$$

where: $s(i,j)$ is the observed covariance between variables i and j,
$r(i,j)$ is the "fitted" covariance predicted by the model for variables i and j,
k is the number of variables, and
Σ_i indicates summation over index i. (3)

The RMSR associated with a model can be thought of as a kind of average difference between observed covariances and covariances predicted by a particular model. The absolute size of the RMSR needs to be intepreted in relation to the magnitude of the observed variances and covariances. However, for a given covariance matrix, differences between the fits of competing models can be evaluated by comparing their RMSR values. Smaller values are associated with better fitting models.

In addition to comparing fit indices, models were evaluated by noting both the pattern and number of large normalized residuals (NR) associated with a particular model. The NR is defined as:

$$NR = [s(i,j)-r(i,j)] / \sqrt{[(s(i,i)s(j,j)+s(i,j)^2)/N]} \qquad (4)$$

where, s and r are as defined above, and N is the number of observational units (in this case, examinees). It should be noted that the numerator in equation (4) is an observed residual value and the denominator is an estimated standard error of that residual. Positive values for NR indicate

that a particular covariance is underfit: the observed covariance is greater than that predicted by the model. Negative values for NR indicate that a particular covariance is overfit: the observed covariance is less than that predicted by the model. Joreskog and Sorbom (1984) suggest that normalized residuals with absolute values greater than two should be examined closely.

RESULTS

Standard Group

Table 1 contains a summary of the fit of the six hypothesized models in the first sample of 1500 standard group students. Listed for each of the models are the chi-square statistic and corresponding degrees of freedom, the RMSR, and the percentage of NRs outside the range of -2.0 to +2.0. The data from standard group sample 1 are not represented well by a 1 factor or 2 factor model. Large differences in both the chi-square and RMSR indices, as well as in the percentage of NRs with absolute values greater than 2, are evident when one compares these models to the either of the four factor models. Of the four factor models, model M4FB, the model in which the structure items form a factor with the writing items, appears to fit substantially better than model M4FA, which has the structure items forming a factor with the reading comprehension items. Despite the fact that both M4FA and M4FB contain the same number of degrees of freedom, the chi-square associated with M4FB is half that associated with M4FA. In addition, both the value of the RMSR and the percentage of NRs with an absolute value greater than 2 are lower for M4FB than they are for M4FA.

An examination of the pattern of NR associated with these models reveals that the inclusion of the language structure items with the reading comprehension in M4FA is clearly not warranted. The NR mean involving the covariance estimates of the three writing items with the three language structure parcels is reduced from 2.415 in M4FA to -.406 in M4FB. Similarly, the NR mean of the covariance estimates for the three language structure parcels with the six reading comprehension parcels changes from -1.496 in M4FA to .208 in M4FB. These changes in the sizes of the normalized residuals, along with the other measures of model fit, demonstrate the

advantage of combining the language structure parcels with the writing items instead of the reading comprehension parcels.

Models M5F and M6F both slightly improve upon the fit of M4FB. The improvements are not trivial, particularly when one considers the changes in chi-square values relative to the associated degrees of freedom. However, the magnitude of the improvements in fit of M5F relative to M4FB, and M6F relative to M5F are clearly less dramatic than those observed with models M1F to M4FB.

Table 2 contains the estimated correlation matrix among the factor scores (with diagonal elements of 1.00 assumed, but omitted, in the table) for models M4FB and M6F for standard group sample 1. It is important to note the correlations among the additional factors included in M6F, relative to the factors included in M4FB. The estimated correlation between the two listening factors is .923. The correlation between the language structure factor and the writing factor is .944. Both of these correlations are noticeably higher than the correlations among the other factors in the model. While both M5F and M6F provide a small improvement in terms of model fit, the high correlations between the two listening factors and between the language structure factor and the writing factor in model M6F could be an indication that the improvement in fit might result from capitalizing on chance factors peculiar to this particular sample.

An additional perspective on possible overfitting can be gained by examining the results of factor analyses involving standard group sample 2. Table 3 presents the fit indices for each of the six models as applied to the covariance matrix derived from the second sample. As in the first sample, models M1F and M2F clearly do not fit the data, and M4FB provides a substantially better fit than does M4FA. However, the addition of a fifth and sixth factor again provides a slight improvement in the fit of the data to the models, particularly with respect to the magnitude of the chi-square statistic.

Table 4 contains estimated correlations among the factor scores for models M4FB and M6F in standard group sample 2. Once again, high correlations between language structure and writing factors (.951) and between the two listening factors (.955) were obtained relative to the other correlations in the table, suggesting an overfitting of the data. However, a

comparison of the correlations between Table 2 and Table 4 reveals two interesting patterns.

First, in both standard group samples, the first listening factor (defined by listening tasks based on short sentences or questions) is more highly correlated with the language structure, writing, and speaking factors than is listening factor 2 (defined by tasks which require comprehension of longer utterances). For example, in standard group sample 1, the estimated correlations of listening factor 1 with the language structure, writing, and speaking factors are .854, .808, and .789, respectively. The corresponding correlations for the listening factor 2 are .802, .717, and .717.

Second, the correlations of the factors measured in the multiple-choice section of the exam (the listening and reading comprehension factors) with the the language structure factor are higher than the corresponding correlations with the writing factor. For example, in Table 2, the correlations of the two listening factors with the language structure factor are .854 and .802, while the corresponding correlations with the writing factor are .808 and .717. The same result can be observed with the data from standard group sample 2 in Table 4.

To summarize, separate analyses based on two independent standard group samples indicate that at least four factors are needed to adequately account for the internal covariation among the sections of the examination. The preferred four factor model is one in which the language structure multiple-choice items are included with writing tasks rather than with the reading comprehension items. However, additional factors, perhaps related to question format, may also be necessary, because improvements in fit for models M5F and M6F were obtained for both samples . While the improvements in fit are small, the correlations between the additional factors included in these models are extremely high. However, the the improvements in model fit for both samples, particularly as measured by the chi-square statistics, for both M5F and M6F are larger than what would be expected solely due to using additional degrees of freedom. Also, the same pattern of estimated factor correlations is observed for both samples.

Standard group 1 and standard group 2 are two samples from the same population. Thus, it is reasonable to expect that the estimates of Γ_x, B, and θ_e obtained from each of these samples are estimates of a common set of

population values. This expectation was tested by fitting both models M4FB and MF6 to both samples simultaneously, while imposing four sets of constraints on the solution. The sets of constraints correspond to four distinct degrees of model invariance that might hold between two different groups (Joreskog and Sorbom, 1984, chapter V). The results of the data analyses of the two samples from standard group provided a baseline to compare the results of similar analyses that involved two different populations.

The first, and most stringent, set of contraints required that a single set of estimates of Γ_x, **B**, and θ_e, be obtained from, and applied to, both sets of data. We refer to this set of contraints as the model of "measurement/structural invariance". The model assumes that the same measurement structure (i.e., **B** and θ_e) holds for both samples. Furthermore, structural invariance implies that the structural relationships among the latent factors (i.e., Γ_x) are invariant across the samples. One might expect the model of measurement/structural-invariance to hold for two samples from the same population, like standard group 1 and standard group 2.

The second, and somewhat less stringent, set of constraints creates what we refer to as a model of "measurement-invariance-only". The model was obtained by removing the restriction of equal Γ_x matrices from the measurement/structural-invariance model. The measurement-invariance-only model allowed the relationships among the latent factors to vary across the samples, but required the measurement properties of the instrument to be the same across populations. Such a model might be expected to hold between nonrandom samples from a single population which differ in level and spread with respect to the factors.

The third set of constraints results in what we refer to as a "invariant-factor-loadings" model. For this model, separate estimates of Γ_x and θ_e are obtained for each sample and only the matrix **B** is constrained to be equal across samples. The fourth set of constraints results in what we refer to as a "invariant-factor-pattern" model. The invariant-factor-pattern model requires only that the pattern of zero and nonzero factor loadings be the same across samples. The values of the nonzero loadings are permitted to be different across samples. Either of these last two models might hold for samples from different populations. The invariant-factor-loadings model

suggests the components of the test measure the underlying factors on the same scales, but with different degrees of precision across populations. The invariant-factor-pattern model suggests that each of the test components measure the same constructs across populations but do so on different scales and with different degrees of precision. This final model is equivalent to fitting a single model (like M4FB), separately for each sample, as was done in Tables 1 and 3.

Table 5 presents a summary of the fit of models M4FB and M6F, for the two standard group samples, with each of the four sets of invariance constraints imposed. As in earlier tables, chi-square statistics (with accompanying degrees of freedom), RMSRs, and NRs are shown for each model. For both M4FB and M6F, the model of measurement/structural-invariance appears to provide a reasonably good fit to the data and is not improved upon substantially by successively relaxing invariance constraints. While the RMSRs are reduced slightly as the sets of constraints are removed, the improvement in fit, as measured by the chi-square values, does not exceed what would be expected, given the concommitant loss in degrees of freedom. Thus, as expected, it is reasonable to assume that the exam exhibits identical measurement properties for both the standard group samples and that the samples exhibit the same level and spread of performance on the factors. Table 6 provides the estimated correlations among the latent factors obtained by fitting models M4FB and M6F simultaneously to both standard group samples, with the measurement/structural-invariance constraints imposed.

Special Group 1

As discussed previously, one purpose of the current research was to compare the internal structure of the exam for different populations of examinees. With this in mind, the same six confirmatory factor analytic models listed in Figure 1 were applied to the data obtained from the 1418 special group 1 students. Table 7 presents the goodness-of-fit indices for all six models. It is evident that the results are quite similar to those obtained with the standard group samples. Once again, model M4FB provides a substantially better fit than does model M4FA. However, one notable difference between the special group 1 analyses and those based on the standard group samples has to do with the relative improvement in fit provided by model M5F over M4FB. The reductions in chi-squares values, RMSR,

and percentage of NRs with absolute values greater than 2 are somewhat larger using the special group 1 data than were the same reductions observed for the standard group samples.

Table 8 shows the estimated correlations among the factors using models M4FB and M6F with the data from special group 1. Results are similar to those reported in Tables 2 and 4. Again, high correlations are found between the language structure and writing factors (.940) and between the two listening factors (.931), relative to the other correlations in the tables. As in the standard group, the first listening factor correlates more strongly with the language structure, writing, and speaking factors than does the the second listening factor. Futhermore, as was found in the standard group, the correlations of the two listening factors and the reading comprehension factor with the language structure factor are larger than the corresponding correlations involving the writing factor.

The four sets of model constraints decribed in the previous section were applied, for both models M4FB and M6F, in order to test for various levels of invariance in the measurement and factor structures between the two standard group samples and special group 1. In conducting multi-sample analyses with both the standard group and special group 1 samples, the measurement/structural-invariance model was imposed with respect to the two standard group samples. Constraints on the equality, across samples, of the various matrices in the models were relaxed only with respect to the special group sample. So, for example, the measurement/structural-invariance model imposes equality constraints on factor loadings, errors of measurement of the observed variables, and relationships among the latent variables for the two standard group samples and the special group 1 population. The model of measurement-invariance-only, allows the value of the Γ_x matrix to differ for the special group 1 sample compared to the combined standard group sample, but constrains the value of Γ_x to be identical across the standard group samples.

Table 9 displays the indices of fit for each of the four sets of invariance constraints for models M4FB and M6F. Unlike the analyses which used only the standard group samples, substantial improvements in fit are obtained when the constraints associated with equal Γ_x and θ_e matrices are relaxed. As a result, invariance cannot be assumed among either the errors

of measurement or the latent factor relationships. Additional improvements in fit occur when the constraint of equal **B** matrices is also relaxed; however, this latter improvement is of somewhat smaller magnitude. This small improvement in fit suggests that at most, only small differences exist in the loadings of the factor pattern matrices.

Special Group 2

Table 10 presents the goodness-of-fit indices for each of the six confirmatory factor analysis models from Figure 1 applied to the 477 special group 2 students. The results are, again, similar to those obtained with the previous three samples. Once again, model M4FB provides a noticeably better fit than does M4FA. As in special group 1, the improvement in fit provided by M5F relative to M4FB is more clear cut than that observed with the standard group samples and an additional, small improvement in fit is generated by M6F.

Table 11 contains the estimated correlations among the factors obtained by applying models M4FB and M6F to the data from special group 2. The pattern of the estimated correlations among the six factors for this population has many similarities to the patterns obtained with the standard groups and special group 1. Again, the correlation between the two listening factors is high (.941); however, the relationship between the language structure and writing factors is somewhat lower (.863) than that observed in the other samples. As in the two previous groups, the correlations between the first listening factor with the the language structure, writing, and speaking factors are higher than the corresponding correlations involving listening factor 2. Futhermore, as in the analyses for the previous three samples, the estimated correlations between the listening factors and the reading comprehension factor with the language structure factor are higher than the corresponding correlations involving the the writing factor.

The same four sets of invariance constraints described in previous section were applied to the data from the two standard groups and special group 2. Table 12 displays the fit of the data to models M4FB and M6F with the various sets of constraints imposed. There is steady and substantial improvement in the fit of both models M4FB and M6F as one moves to less restrictive sets of invariance constraints. Invariance cannot be assumed

among either the errors of measurement, latent factor relationships, or the factor loadings between the standard group and special group 2 samples.

College Standard Group

Table 13 presents the goodness-of-fit indices for each of the six factor analysis models from Figure 1 applied to the 302 standard group college students. Results are remarkably consistent with those obtained with the previous samples, given that the administration conditions for the college samples were somewhat different from those associated with the AP samples (see Morgan & Maneckshana, 1988). Again, model M4FB fits better than M4FA. Model M5F results in an improvement in fit relative to M4FB. An additional, albeit smaller, improvement is obtained for model M6F relative to M5F.

Table 14 displays the estimated correlations among the factors obtained by fitting models M4FB and M6F to the data from the college sample. Overall, the correlations among the factors are generally lower for this group than for the AP sample groups. For example, the correlation between language structure and writing is considerably lower than in the other samples (.773). However, the correlations between the two listening factors remains high (.920). In general, the other patterns observed with the previous samples are present in Table 14.

Again, the same four sets of invariance constraints were applied to the data from the college sample and the two standard group samples. Table 15 displays the fit of the data for each of the four models. Comparisons of the percent of large NRs, RMSR, and the Chi-Squares for the first two models suggests that the assumption of invariant factor matrices (Γ_x) between the standard high school group and the standard college group is violated. This is consistent with the differences in the factor correlations for these two groups found in Tables 6 and 14. A comparison of models two and three shows trivial differences in the percentage of large NRs and the RMSR and a small difference, relative to degrees of freedom, in the values of the chi-squares. Consequently, the model of measurement-invariance-only does not seem to be improved upon by further relaxing of constraints.

DISCUSSION

One of the stated purposes of the present research was to compare the internal covariance structure of the AP French Language exam with prior notions as to the kinds of skills and knowledge which are intended to be measured. Results of separate factor analyses carried out on data from samples from four different test-taking populations are consistent in suggesting that at least four latent dimensions are required to adequately model the covariance among the sets of items in the test. One reasonable interpretation of these dimensions, given the sets of items which are permitted to load on these dimensions, is that they correspond to examinee proficency with respect to listening, reading, writing, and speaking French.

However, the results also suggest that prior notions as to the proficiencies measured by the set of structure multiple-choice items are likely wrong. It had been thought that the structure items should be grouped with the reading comprehension items as measures of reading proficiency. It appears that, at the most, the structure items measure a dimension similar to that which is measured by the short-answer and free-response writing tasks. As a practical result of this finding, the weighting scheme of the test, which assumes a four factor model where language structure and reading comprehension form a single factor, needs to be modified.

The intention of the test developers is to assign equal weight to the four basic language skills. The current test design results in an overweighting of the tasks strongly related to identifying and producing grammatically correct prose, while underweighting the ability to read and interpret French passages. One solution could be to simply change to weighting of the subsections to be congruent with both the intended and actual underlying structure of the exam. Another possibility would be to eliminate the language structure section, and replace it by an expansion of the reading comprehension and/or writing sections.

Analyses across all four samples also suggest that the separate factors associated with the language structure and listening tasks do capture something relevant about the underlying structure of the test, rather than peculiarities associated with particular samples. A consistent pattern of interrelationships is present among the factors in the six-factor solution

across all four populations. What is not clear from the analyses is the utility of moving from a more parsimonious four factor representation to the more complicated six factor representation of the exam.

The language structure factor (which is measured by multiple-choice parcels) correlates higher with those factors which are also measured by parcels of multiple-choice items than does the writing factor, which is measured with free-response questions. The writing factor, in turn, correlates more highly with the free-response format speaking factor than does the language structure factor. However, the language structure and writing factors are highly correlated, particularly in the standard group and the special group 1 samples.

It is interesting to note that the grading of the writing tasks is based in part on correct grammatical usage. One might speculate that, with the exception of those students with extensive out-of-school French language experience, the ability to produce grammatically correct prose is being measured in both a multiple-choice and a free-response framework. As a result, language structure and writing tasks might be measuring the same constructs, but with slightly different item types.

In all four samples, analyses also indicated that including two separate, but highly correlated, listening factors improved the fit of the models slightly. One difference that consistantly appears is that the first listening factor correlates higher with the other factors, with the exception of the reading comprehension factor, than does the second listening factor. One might speculate that this pattern of correlations may result from the long listening and reading comprehension passages item types uniquely tapping the ability to retain in memory long spoken or written passages, as well as measuring abilities specific to knowledge of the French language.

The second major purpose of the current research was to compare the covariance structures of the exam across different populations of examinees. The confirmatory factor analyses, with sets of equality constraints imposed, appear to indicate a large degree of invariance in the structure of the data for the college and high school standard groups. Indeed, a model which assumes equal factor loadings and error variances for high school and college standard group examinees appears to provide an adequate fit to the data.

A lesser degree of congruence is observed between the standard and special high school populations. In the multi-sample analyses, the more out-of-class French language experience that a population demonstrated, the more dissimilar were the measurement properties of the test and factor interrelationships for that group compared to those properties and interrelationships observed for the standard group. This result would seem to demonstrate that testing and learning circumstances (high school vs. college) have less effect upon the dimensional structure of the test than population characteristics (as measured by out-of-class French experience).

The parcel method proved useful in this analysis, however, the amount of variation in factor loadings and factor correlations that might be observed if the item composition of the parcels were different is not known. In the course of this study, however, a second set of data analyses was conducted using different reading comprehension and listening parcels than those reported above. In this second set of analyses, parcels were formed by grouping the items which were based on the same reading or listening passages. This second set of analyses was not reported since little difference existed between it and the analyses discussed above. While this finding does not provide anything close to a thorough test of the effects of item parcel composition on the factor analytic results, it does provide some evidence that the results may not be strongly dependent on the compostion of the parcels.

In summary, confirmatory factor analyses of the French Language exam yielded the following conclusions: 1) The exam most likely measures at least four major dimensions which we have associated with the language proficiencies of listening, reading, writing, and speaking; 2) Two additional dimensions may be present which appear to be related to the aspects of item format; 3) For both high school and college standard group samples, the exam appears to measure the same constructs, on the same scale, with the same degree of precision; 4) For the different special group populations, the exam appears to measure the same constructs, on slightly different scales, with differing degrees of measurement precision.

References

Bock, R.D., Gibbons, R.D., & Muraki, E. (1986). Full information item factor analysis (MRC Report No. 84-1 revised). Chicago, IL: National Opinion Research Center.

Carroll, J.B. (1983). The difficulty of a test and its factor composition revisited. In S. Messick and H. Wainer (Eds.), Principals of modern psychological measurement: A festschrift for Frederick M. Lord. Hillsdale, NJ: Erlbaum.

College Board. (1987). Advanced placement course description: French. New York: College Entrance Examination Board.

Cook, L.L., Dorans, N.J., & Eignor, D.R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. Journal of Educational Statistics, 13, 19-43.

Dorans, N.J. & Lawrence, I.M. (1987). The internal construct validity of the SAT (ETS Research Report No. RR-87-35). Princeton, N.J.: Educational Testing Service.

Joreskog, K.G. & Sorbom, D. (1984). LISREL VI - Analysis of linear structural relationships by the method of maximum likelihood. Chicago, IL: International Educational Services.

Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 9, 139-164.

Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Morgan, R.L. & Maneckshana, B. (1988). The 1987 advanced placement college comparability study in French Language (Statistical Report No. SR-88-09). Princeton, N.J.: Educational Testing Service.

Wilson, D., Wood, R.L., & Gibbons, R. (1984). TESTFACT: Test scoring and item factor analysis. Chicago, IL: Scientific Software.

Zwick, R. (1986). Assessment of the dimensionality of NAEP year 15 reading data (ETS Research Report No. RR-86-4). Princeton, N.J.: Educational Testing Service.

TABLE 1

Indices of Model Fit
Standard Group - Sample 1

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
M1F	1749.64 (170)	1.339	24.2
M2F	1472.21 (169)	1.294	22.6
M4FA	704.28 (164)	.958	11.1
M4FB	336.87 (164)	.558	2.1
M5F	288.62 (160)	.500	1.6
M6F	235.00 (155)	.433	0.5

TABLE 2

Correlations Among Factors
Standard Group - Sample 1

Model M4FB

	Writing	Reading Comp.	Speaking
Listening	.829	.790	.778
Writing		.820	.779
Reading Comp.			.631

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.923	.854	.772	.808	.789
Listening 2		.802	.774	.717	.717
Language Struc.			.818	.944	.746
Reading Comp.				.799	.631
Writing					.780

TABLE 3

Indices of Model Fit
Standard Group - Sample 2

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
M1F	1736.48 (170)	1.205	21.6
M2F	1399.25 (169)	1.145	18.4
M4FA	732.05 (164)	.925	10.0
M4FB	296.91 (164)	.532	1.1
M5F	241.06 (160)	.444	0.5
M6F	223.43 (155)	.415	0.5

TABLE 4

Correlations Among Factors
Standard Group - Sample 2

Model M4FB

	Writing	Reading Comp.	Speaking
Listening	.844	.819	.740
Writing		.791	.746
Reading Comp.			.646

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.954	.885	.806	.812	.742
Listening 2		.840	.812	.765	.712
Language Struc.			.811	.951	.722
Reading Comp.				.762	.646
Writing					.744

TABLE 5

Indices of Model Fit
Multi-Sample Analyses
Standard Group - Samples 1&2

Model M4FB

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	686.81 (374)	.641	1.7
Measurement Invariance	665.02 (364)	.623	1.7
Factor Loading Invariance	648.06 (344)	.622	1.7
Factor Pattern Invariance	633.78 (328)	.545	1.6

Model M6F

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	519.98 (365)	.547	0.2
Measurement Invariance	489.00 (344)	.507	0.2
Factor Pattern Invariance	471.15 (324)	.508	0.2
Factor Pattern Invariance	458.43 (310)	.424	0.5

TABLE 6

Correlations Among Factors
Standard Group - Invariant Solution

Model M4FB

	Writing	Reading Comp.	Speaking
Listening	.836	.805	.754
Writing		.806	.762
Reading Comp.			.639

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.938	.868	.789	.809	.765
Listening 2		.821	.793	.741	.714
Language Struc.			.815	.947	.733
Reading Comp.				.780	.639
Writing					.761

TABLE 7

Indices of Model Fit
Special Group 1

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
M1F	2258.32 (170)	1.286	25.8
M2F	1854.29 (169)	1.228	26.8
M4FA	929.51 (164)	1.040	19.5
M4FB	508.47 (164)	.672	6.3
M5F	368.65 (160)	.520	1.6
M6F	327.44 (155)	.483	1.6

TABLE 8

Correlations Among Factors
Special Group 1

Model M4FB

	Grammar	Reading Comp.	Speaking
Listening	.848	.769	.779
Grammar		.787	.781
Reading Comp.			.540

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.940	.912	.752	.778	.790
Listening 2		.872	.760	.720	.728
Language Struc.			.823	.931	.788
Reading Comp.				.727	.540
Writing					.746

TABLE 9

Indices of Model Fit
Multi-Sample Analyses
Invariant AP Standard Group with Special Group 1

Model M4FB

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	1645.40 (584)	1.269	16.2
Measurement Invariance	1558.93 (574)	.840	4.4
Factor Loading Invariance	1241.42 (554)	.740	3.8
Factor Pattern Invariance	1195.28 (538)	.651	3.1

Model M6F

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	1335.02 (575)	1.212	14.3
Measurement Invariance	1173.53 (554)	.713	3.2
Factor Loading Invariance	895.46 (534)	.628	2.1
Factor Pattern Invariance	847.42 (520)	.527	0.9

TABLE 10

Indices of Model Fit
Special Group 2

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
M1F	940.13 (170)	1.224	8.9
M2F	834.54 (169)	1.195	6.8
M4FA	585.17 (164)	1.118	3.2
M4FB	407.34 (164)	.704	1.6
M5F	309.60 (160)	.652	0.5
M6F	270.41 (155)	.599	0.5

TABLE 11

Correlations Among Factors
Special Group 2

Model M4FB

	Writing	Reading Comp.	Speaking
Listening	.911	.802	.860
Writing		.810	.850
Reading Comp.			.665

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.941	.929	.779	.854	.884
Listening 2		.852	.811	.740	.777
Language Struc.			.776	.863	.829
Reading Comp.				.775	.666
Writing					.799

TABLE 12

Indices of Model Fit
Multi-Sample Analyses
Invariant AP Standard Group with Special Group 2

Model M4FB

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	1991.08 (584)	2.084	13.5
Measurement Invariance	1852.37 (574)	1.324	7.5
Factor Loading Invariance	1188.46 (554)	1.014	6.2
Factor Pattern Invariance	1094.15 (538)	.663	1.6

Model M6F

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	1752.70 (575)	2.055	12.5
Measurement Invariance	1490.55 (554)	1.201	5.4
Factor Loading Invariance	888.16 (534)	.926	4.6
Factor Pattern Invariance	790.39 (520)	.564	0.4

TABLE 13

Indices of Model Fit
College - Standard Group

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
M1F	759.01 (170)	2.449	16.3
M2F	617.96 (169)	2.357	14.7
M4FA	357.65 (164)	2.017	7.4
M4FB	274.58 (164)	1.379	3.7
M5F	178.45 (160)	.986	0.0
M6F	146.57 (155)	.854	0.0

TABLE 14

Correlations Among Factors
College - Standard Group

Model M4FB

	Writing	Reading Comp.	Speaking
Listening	.714	.618	.512
Writing		.720	.674
Reading Comp.			.527

Model M6F

	List. 2	Lang. Str.	Read. Comp.	Writ.	Speak.
Listening 1	.920	.820	.654	.634	.547
Listening 2		.721	.531	.400	.434
Language Struc.			.632	.773	.527
Reading Comp.				.696	.525
Writing					.704

TABLE 15

Indices of Model Fit
Multi-Sample Analyses
Invariant AP Standard Group with College Standard Group

Model M4FB

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	1086.17 (584)	1.602	13.0
Measurement Invariance	1013.17 (574)	1.241	3.7
Factor Loading Invariance	977.09 (554)	1.190	2.9
Factor Pattern Invariance	961.39 (538)	.953	3.0

Model M6F

Model	Chi-Square (df)	Root Mean Square Residual	% NR > |2|
Measurement/ Structural Invariance	835.26 (575)	1.554	10.8
Measurement Invariance	718.28 (554)	.898	0.3
Factor Loading Invariance	677.45 (534)	.819	0.3
Factor Pattern Invariance	666.55 (520)	.719	0.3

FIGURE 1

Hypothesized Factor Pattern Models

	Model					
Parcel	M1F	M2F	M4FA	M4FB	M5F	M6F
L1-1	X	X0	X000	X000	X0000	X00000
L1-2	X	X0	X000	X000	X0000	X00000
L1-3	X	X0	X000	X000	X0000	X00000
L2-1	X	X0	X000	X000	X0000	0X0000
L2-2	X	X0	X000	X000	X0000	0X0000
L2-3	X	X0	X000	X000	X0000	0X0000
LS-1	X	X0	0X00	0X00	0X000	00X000
LS-2	X	X0	0X00	0X00	0X000	00X000
LS-3	X	X0	0X00	0X00	0X000	00X000
RC-1	X	X0	0X00	00X0	00X00	000X00
RC-2	X	X0	0X00	00X0	00X00	000X00
RC-3	X	X0	0X00	00X0	00X00	000X00
RC-4	X	X0	0X00	00X0	00X00	000X00
RC-5	X	X0	0X00	00X0	00X00	000X00
RC-6	X	X0	0X00	00X0	00X00	000X00
WR-1	X	0X	00X0	0X00	000X0	0000X0
WR-2	X	0X	00X0	0X00	000X0	0000X0
WR-3	X	0X	00X0	0X00	000X0	0000X0
SP-1	X	0X	000X	000X	0000X	00000X
SP-2	X	0X	000X	000X	0000X	00000X

X = Factor Loading
0 = Loading Fixed to Zero
L1 = Short Listening Task
L2 = Long Listening Task
LS = Language Structure
RC = Reading Comprehension
WR = Writing
SP = Speaking

APPENDIX 1

Colleges Providing Data

Boston University

California State University at Los Angeles

Catholic University (D.C.)

Furman University

Georgetown University

Holy Cross University

Manhattan College

Middlebury College

Mundelein College

Oregon State University

Portland State University

St. Lawrence University

University of Arizona

University of Virginia

University of Wisconsin - Madison

Wake Forest University

APPENDIX 2

Mean and Standard Deviation (As Percent of Maximum)
for Each Parcel in Each Sample

Parcel	Standard 1		Standard 2		Special 1		Special 2		College	
	M	SD	M	SD	M	SD	M	SD	M	SD
L1-1	58	22	57	22	75	21	87	17	52	23
L1-2	58	22	57	22	73	21	85	19	54	22
L1-3	56	25	56	25	75	23	88	18	51	25
L2-1	58	23	56	23	73	21	82	18	55	23
L2-2	58	22	57	22	74	21	84	19	56	24
L2-3	57	23	56	23	75	22	83	19	54	24
LS-1	53	25	53	24	67	24	80	23	49	23
LS-2	54	23	53	23	64	22	75	23	49	24
LS-3	54	25	54	24	66	23	77	22	46	23
RC-1	71	21	71	21	79	19	82	18	70	21
RC-2	71	23	71	23	78	20	82	20	67	23
RC-3	71	22	71	22	79	19	81	18	70	21
RC-4	72	22	71	22	79	19	81	19	72	20
RC-5	71	21	72	21	77	20	81	20	71	21
RC-6	72	21	71	22	78	20	79	19	70	23
WR-1	54	20	53	21	64	21	75	21	51	19
WR-2	58	26	58	26	64	25	67	24	49	22
WR-3	51	17	50	17	59	19	68	23	50	16
SP-1	63	16	62	16	74	15	86	15	61	14
SP-2	57	17	57	18	72	18	86	18	56	15

L1 = Short Listening Task
L2 = Long Listening Task
LS = Language Structure
RC = Reading Comprehension
WR = Writing
SP = Speaking

APPENDIX 3

CORRELATION MATRIX - STANDARD GROUP - SAMPLE 1

	L1-1	L1-2	L1-3	L2-1	L2-2	L2-3	LS-1	LS-2	LS-3	RC-1
L1-1	1.000									
L1-2	0.567	1.000								
L1-3	0.564	0.557	1.000							
L2-1	0.495	0.482	0.507	1.000						
L2-2	0.476	0.437	0.474	0.454	1.000					
L2-3	0.474	0.445	0.437	0.467	0.466	1.000				
LS-1	0.524	0.488	0.506	0.456	0.428	0.426	1.000			
LS-2	0.423	0.430	0.396	0.372	0.325	0.335	0.501	1.000		
LS-3	0.405	0.415	0.389	0.354	0.336	0.340	0.496	0.455	1.000	
RC-1	0.431	0.401	0.400	0.410	0.364	0.386	0.423	0.388	0.389	1.000
RC-2	0.471	0.447	0.476	0.427	0.426	0.416	0.492	0.424	0.458	0.564
RC-3	0.468	0.435	0.439	0.431	0.375	0.414	0.484	0.406	0.420	0.537
RC-4	0.435	0.412	0.429	0.402	0.381	0.356	0.454	0.411	0.422	0.530
RC-5	0.370	0.361	0.349	0.343	0.309	0.350	0.441	0.348	0.347	0.500
RC-6	0.410	0.425	0.397	0.375	0.347	0.346	0.400	0.374	0.357	0.484
WR-1	0.507	0.491	0.506	0.428	0.388	0.412	0.581	0.521	0.502	0.459
WR-2	0.437	0.440	0.418	0.349	0.309	0.337	0.526	0.494	0.496	0.402
WR-3	0.471	0.455	0.463	0.379	0.390	0.368	0.529	0.452	0.442	0.381
SP-1	0.469	0.458	0.492	0.394	0.406	0.381	0.478	0.413	0.378	0.339
SP-2	0.497	0.468	0.516	0.411	0.402	0.395	0.483	0.391	0.382	0.341

	RC-2	RC-3	RC-4	RC-5	RC-6	WR-1	WR-2	WR-3	SP-1	SP-2
RC-2	1.000									
RC-3	0.628	1.000								
RC-4	0.587	0.554	1.000							
RC-5	0.528	0.520	0.472	1.000						
RC-6	0.532	0.497	0.518	0.449	1.000					
WR-1	0.525	0.511	0.495	0.425	0.446	1.000				
WR-2	0.495	0.461	0.446	0.391	0.393	0.614	1.000			
WR-3	0.487	0.448	0.430	0.383	0.392	0.570	0.591	1.000		
SP-1	0.430	0.395	0.424	0.322	0.343	0.519	0.479	0.478	1.000	
SP-2	0.430	0.392	0.403	0.306	0.329	0.494	0.466	0.508	0.668	1.000

BEST COPY AVAILABLE

CORRELATION MATRIX - STANDARD GROUP - SAMPLE 2

	L1-1	L1-2	L1-3	L2-1	L2-2	L2-3	LS-1	LS-2	LS-3	RC-1
L1-1	1.000									
L1-2	0.518	1.000								
L1-3	0.558	0.536	1.000							
L2-1	0.515	0.444	0.504	1.000						
L2-2	0.456	0.460	0.497	0.440	1.000					
L2-3	0.438	0.408	0.438	0.443	0.439	1.000				
LS-1	0.493	0.459	0.501	0.452	0.428	0.407	1.000			
LS-2	0.441	0.406	0.430	0.382	0.352	0.340	0.493	1.000		
LS-3	0.457	0.415	0.451	0.374	0.387	0.345	0.490	0.468	1.000	
RC-1	0.436	0.375	0.404	0.395	0.402	0.363	0.423	0.378	0.363	1.000
RC-2	0.496	0.474	0.475	0.456	0.422	0.402	0.478	0.469	0.432	0.565
RC-3	0.482	0.445	0.469	0.436	0.424	0.402	0.474	0.413	0.397	0.511
RC-4	0.468	0.422	0.460	0.416	0.402	0.397	0.444	0.410	0.391	0.529
RC-5	0.371	0.338	0.371	0.352	0.347	0.353	0.407	0.359	0.359	0.473
RC-6	0.455	0.373	0.426	0.387	0.378	0.366	0.442	0.404	0.362	0.493
WR-1	0.513	0.462	0.494	0.435	0.429	0.440	0.565	0.502	0.520	0.433
WR-2	0.451	0.419	0.471	0.376	0.346	0.352	0.563	0.536	0.525	0.401
WR-3	0.483	0.438	0.490	0.417	0.417	0.395	0.521	0.491	0.476	0.438
SP-1	0.463	0.437	0.466	0.421	0.383	0.364	0.443	0.388	0.400	0.381
SP-2	0.448	0.434	0.468	0.408	0.384	0.395	0.459	0.397	0.410	0.391

	RC-2	RC-3	RC-4	RC-5	RC-6	WR-1	WR-2	WR-3	SP-1	SP-2
RC-2	1.000									
RC-3	0.606	1.000								
RC-4	0.606	0.547	1.000							
RC-5	0.556	0.510	0.496	1.000						
RC-6	0.592	0.521	0.518	0.495	1.000					
WR-1	0.502	0.484	0.474	0.432	0.457	1.000				
WR-2	0.480	0.438	0.450	0.387	0.418	0.632	1.000			
WR-3	0.467	0.418	0.447	0.367	0.416	0.619	0.619	1.000		
SP-1	0.422	0.410	0.422	0.317	0.377	0.490	0.464	0.507	1.000	
SP-2	0.422	0.430	0.414	0.330	0.399	0.490	0.471	0.515	0.692	1.000

CORRELATION MATRIX - SPECIAL GROUP 1

	L1-1	L1-2	L1-3	L2-1	L2-2	L2-3	LS-1	LS-2	LS-3	RC-1
L1-1	1.000									
L1-2	0.621	1.000								
L1-3	0.602	0.605	1.000							
L2-1	0.558	0.523	0.543	1.000						
L2-2	0.524	0.525	0.513	0.522	1.000					
L2-3	0.554	0.532	0.523	0.502	0.557	1.000				
LS-1	0.610	0.567	0.584	0.513	0.497	0.529	1.000			
LS-2	0.469	0.431	0.405	0.374	0.372	0.401	0.497	1.000		
LS-3	0.499	0.482	0.480	0.460	0.441	0.452	0.549	0.499	1.000	
RC-1	0.406	0.390	0.406	0.405	0.390	0.385	0.456	0.348	0.414	1.000
RC-2	0.443	0.476	0.451	0.418	0.431	0.437	0.492	0.418	0.458	0.552
RC-3	0.464	0.455	0.430	0.407	0.409	0.414	0.485	0.442	0.455	0.502
RC-4	0.449	0.451	0.459	0.406	0.412	0.414	0.477	0.408	0.452	0.534
RC-5	0.375	0.360	0.363	0.356	0.337	0.348	0.388	0.356	0.363	0.459
RC-6	0.389	0.413	0.360	0.389	0.378	0.359	0.435	0.350	0.386	0.469
WR-1	0.560	0.550	0.534	0.456	0.460	0.467	0.610	0.536	0.567	0.427
WR-2	0.390	0.403	0.364	0.347	0.344	0.345	0.516	0.491	0.473	0.378
WR-3	0.456	0.466	0.443	0.406	0.402	0.383	0.530	0.501	0.478	0.375
SP-1	0.531	0.494	0.499	0.454	0.442	0.425	0.533	0.447	0.466	0.295
SP-2	0.558	0.554	0.527	0.481	0.461	0.454	0.558	0.436	0.450	0.305

	RC-2	RC-3	RC-4	RC-5	RC-6	WR-1	WR-2	WR-3	SP-1	SP-2
RC-2	1.000									
RC-3	0.585	1.000								
RC-4	0.583	0.513	1.000							
RC-5	0.520	0.421	0.485	1.000						
RC-6	0.553	0.467	0.519	0.456	1.000					
WR-1	0.472	0.455	0.435	0.396	0.423	1.000				
WR-2	0.432	0.412	0.414	0.341	0.381	0.617	1.000			
WR-3	0.427	0.415	0.419	0.331	0.373	0.627	0.585	1.000		
SP-1	0.363	0.346	0.358	0.288	0.284	0.539	0.438	0.517	1.000	
SP-2	0.363	0.360	0.363	0.312	0.314	0.556	0.406	0.514	0.734	1.000

CORRELATION MATRIX - SPECIAL GROUP 2

	L1-1	L1-2	L1-3	L2-1	L2-2	L2-3	LS-1	LS-2	LS-3	RC-1
L1-1	1.000									
L1-2	0.668	1.000								
L1-3	0.608	0.641	1.000							
L2-1	0.518	0.508	0.542	1.000						
L2-2	0.572	0.565	0.600	0.518	1.000					
L2-3	0.534	0.513	0.504	0.450	0.549	1.000				
LS-1	0.656	0.665	0.622	0.535	0.585	0.482	1.000			
LS-2	0.581	0.620	0.607	0.493	0.528	0.412	0.701	1.000		
LS-3	0.541	0.554	0.579	0.479	0.504	0.404	0.640	0.645	1.000	
RC-1	0.487	0.491	0.418	0.399	0.488	0.395	0.482	0.498	0.435	1.000
RC-2	0.557	0.543	0.501	0.480	0.499	0.399	0.550	0.537	0.499	0.598
RC-3	0.473	0.489	0.436	0.473	0.521	0.461	0.517	0.511	0.474	0.536
RC-4	0.522	0.449	0.399	0.392	0.511	0.374	0.477	0.463	0.416	0.573
RC-5	0.427	0.436	0.353	0.345	0.423	0.307	0.434	0.483	0.368	0.541
RC-6	0.516	0.425	0.390	0.394	0.451	0.370	0.461	0.419	0.399	0.494
WR-1	0.617	0.660	0.612	0.478	0.557	0.455	0.661	0.644	0.616	0.550
WR-2	0.422	0.506	0.477	0.336	0.421	0.283	0.507	0.535	0.512	0.395
WR-3	0.495	0.562	0.477	0.381	0.426	0.310	0.535	0.549	0.480	0.439
SP-1	0.607	0.672	0.638	0.512	0.566	0.439	0.641	0.639	0.562	0.463
SP-2	0.590	0.634	0.622	0.490	0.513	0.410	0.618	0.583	0.535	0.484

	RC-2	RC-3	RC-4	RC-5	RC-6	WR-1	WR-2	WR-3	SP-1	SP-2
RC-2	1.000									
RC-3	0.588	1.000								
RC-4	0.590	0.594	1.000							
RC-5	0.577	0.502	0.541	1.000						
RC-6	0.559	0.522	0.598	0.523	1.000					
WR-1	0.616	0.520	0.503	0.448	0.481	1.000				
WR-2	0.472	0.427	0.393	0.406	0.416	0.680	1.000			
WR-3	0.496	0.426	0.397	0.396	0.466	0.641	0.650	1.000		
SP-1	0.529	0.452	0.397	0.412	0.402	0.647	0.514	0.575	1.000	
SP-2	0.505	0.454	0.407	0.389	0.374	0.639	0.478	0.540	0.788	1.000

BEST COPY AVAILABLE

CORRELATION MATRIX - COLLEGE GROUP

	L1-1	L1-2	L1-3	L2-1	L2-2	L2-3	LS-1	LS-2	LS-3	RC-1
L1-1	1.000									
L1-2	0.575	1.000								
L1-3	0.578	0.587	1.000							
L2-1	0.518	0.544	0.579	1.000						
L2-2	0.493	0.539	0.521	0.523	1.000					
L2-3	0.440	0.459	0.481	0.555	0.521	1.000				
LS-1	0.479	0.470	0.495	0.440	0.409	0.424	1.000			
LS-2	0.438	0.410	0.431	0.408	0.344	0.316	0.474	1.000		
LS-3	0.440	0.466	0.397	0.337	0.415	0.272	0.564	0.484	1.000	
RC-1	0.306	0.374	0.380	0.297	0.367	0.254	0.317	0.311	0.312	1.000
RC-2	0.426	0.417	0.390	0.279	0.355	0.268	0.418	0.345	0.379	0.593
RC-3	0.378	0.424	0.374	0.322	0.354	0.229	0.357	0.343	0.352	0.529
RC-4	0.307	0.327	0.328	0.308	0.270	0.227	0.327	0.251	0.322	0.510
RC-5	0.332	0.298	0.308	0.231	0.228	0.239	0.288	0.257	0.308	0.454
RC-6	0.310	0.332	0.266	0.190	0.321	0.164	0.252	0.264	0.293	0.500
WR-1	0.336	0.338	0.362	0.207	0.252	0.143	0.442	0.339	0.465	0.276
WR-2	0.385	0.363	0.387	0.212	0.242	0.186	0.489	0.413	0.435	0.346
WR-3	0.369	0.358	0.390	0.280	0.291	0.191	0.397	0.342	0.381	0.386
SP-1	0.260	0.324	0.318	0.240	0.185	0.202	0.342	0.170	0.284	0.271
SP-2	0.356	0.310	0.365	0.292	0.250	0.277	0.351	0.284	0.283	0.289

	RC-2	RC-3	RC-4	RC-5	RC-6	WR-1	WR-2	WR-3	SP-1	SP-2
RC-2	1.000									
RC-3	0.586	1.000								
RC-4	0.549	0.475	1.000							
RC-5	0.514	0.426	0.436	1.000						
RC-6	0.512	0.482	0.451	0.424	1.000					
WR-1	0.393	0.404	0.314	0.320	0.308	1.000				
WR-2	0.462	0.406	0.381	0.402	0.358	0.600	1.000			
WR-3	0.433	0.425	0.350	0.363	0.344	0.508	0.568	1.000		
SP-1	0.332	0.306	0.379	0.232	0.271	0.388	0.394	0.398	1.000	
SP-2	0.307	0.303	0.287	0.307	0.249	0.426	0.463	0.407	0.601	1.000